# Survey Evaluation and Data Quality Activities in the Energy Information Administration (EIA)

Nancy Kirkendall and Renee Miller
EIA

## ABSTRACT

In the Energy Information Administration (EIA) the responsibility for the evaluation and maintenance of the accuracy of published data falls in two places: the specific program office which collects and publishes the data and the Office of Statistical Standards which has general oversight and evaluation responsibility. In this paper we will present examples of both types of activities. First, the Petroleum Supply Division's program for verifying and maintaining the accuracy of their data will be described. The Petroleum Supply Division collects data on the supply and movement of petroleum in the United States on both their weekly sample surveys and on their monthly census surveys. This reporting procedure provides a unique opportunity for verifying and evaluating published data. Second, the survey evaluations performed by the Office of Statistical Standards will be described. These analyses are based on two approaches: to compare EIA data with other similar data series; and to examine the structure of the EIA data series for anomalies.

## INTRODUCTION

The Energy Information Administration (EIA) is organized into three program offices: the Office of Oil and Gas, the Office of Coal, Nuclear, Electric and Alternate Fuels, and the Office of Energy Markets and End Use; and four support offices: the National Energy Information Center, the Office of Automatic Data Processing Services, the Office of Planning and Resources, and the Office of Statistical Standards. The responsibility for the quality of data collected and published by EIA resides in the three program offices, which operate the surveys and publish the data, and in the Office of Statistical Standards which is responsible for the overall quality of EIA data.

The program offices have the final responsibility for the accuracy of their data: including the design of the survey forms and sampling strategy, maintenance of frames, design, and implementation of the survey processing systems (including edit and imputation procedures), and the publication of the data. It is the program office's responsibility to maintain the quality of each phase of its operation.

The Office of Statistical Standards (OSS) serves as a statistical advisor to the program offices, and has a general oversight function. The oversight function is accomplished by several diverse activities:

1. All survey clearance packages to be sent to the Office of Management and Budget, new publications, and revisions to ongoing publications are reviewed by OSS. These reviews assure conformance with EIA Standards, and identify problems which require further work. In special cases these reviews are performed by experts outside of EIA under OSS auspices.

2. OSS manages the Quality Audit program in which individual survey systems and operations are subjected to a detailed scrutiny, from the receipt and logging in of the survey forms to the computer processing and publication of the data. As a result of the audit, recommendations for improvements to the system are made and (usually) agreed to by the program offices. Almost half of the survey systems in EIA have received a quality audit within the last 3 years.

3. OSS is responsible for formulating standards which dictate documentation requirements for models and systems, specify conventions for graphs, prescribe disclosure avoidance procedures, specify maintenance of performance statistics, etc. These standards assure maintenance of good practice and uniformity of methodology, documentation, and presentation in EIA products. They are designed to alleviate problems discovered in the review processes and are often based on OSS research. The standards are agreed to by the other offices within EIA.

This is a working document prepared in support of the Census Research Conference, it is not an official Energy Information Administration publication.

4. Each year OSS performs an in-depth analysis of a data series, or a collection of data series, produced by one of the program offices. The analysis includes a comparison of the published EIA data with data from other sources, where possible, and/or an identification and evaluation of internal sources of variance such as outliers, inconsistencies, reporting error, processing error, etc.

This paper gives examples of survey evaluation and data quality activities in EIA. The first example will summarize the work done within one of the program offices: the Petroleum Supply Division (PSD) of the Office of Oil and Gas. The second example will describe the methods and results of the OSS analysis of data series (the State-of-the Data Reports).

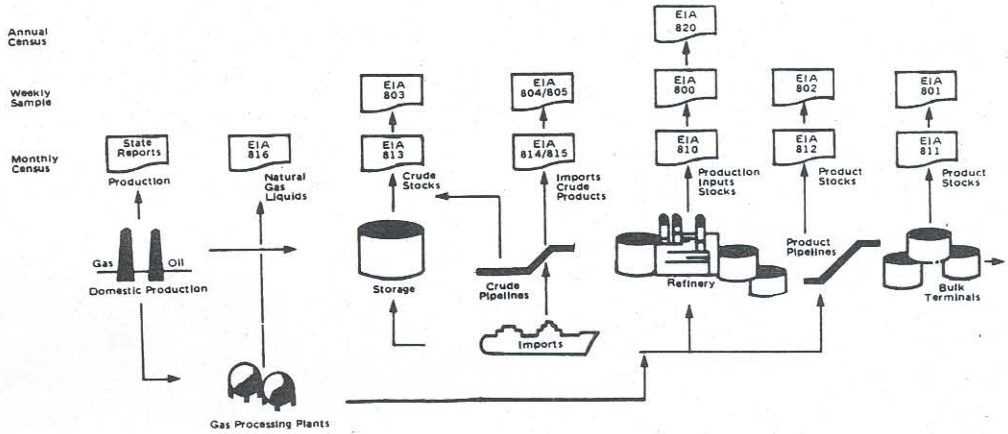
## DATA QUALITY IN THE PETROLEUM SUPPLY DIVISION

### An Overview of Petroleum Supply Data

The Petroleum Supply Division collects and publishes statistics concerning supply and movement of petroleum in the United States. The surveys contributing data to the Petroleum Supply Reporting System (PSRS) measure supply and throughput at various points in the petroleum supply flow, from the production of crude oil to the distribution of petroleum products. EIA publishes data from the PSRS in the *Weekly Petroleum Status Report* (*WPSR*), the *Petroleum Supply Monthly* (*PSM*), the *Petroleum Supply Annual* (*PSA*), the *Monthly Energy Review* (*MER*), and the *Annual Energy Review* (*AER*).

Petroleum supply incorporates domestic production, foreign trade, refinery operations, stocks, and transportation. Surveys collect data weekly, monthly, and annually. The relationship between the PSRS and the petroleum supply network is illustrated in Figure 1.

Figure 1. The Petroleum Supply Reporting System



The primary focus of this presentation will be petroleum inventories and production reported on the monthly and weekly survey forms. The frames for these forms are crude stockholders with stocks of more than 1,000 barrels, refiners, product pipelines, and bulk terminals which have more than 50 thousand barrels capacity and/or receive products by tanker, barge or pipeline. These forms provide data on the weekly and monthly inventories of crude oil, inputs to refineries, and production and inventories of petroleum products.

The monthly surveys are all census surveys. Data from these surveys are published in the *PSM* 2 months after the close of the report month. The weekly surveys are sample surveys. Data from the weekly surveys are published in the *WPSR* less than 1 week after the close of the report week. The primary petroleum supply populations are relatively small and highly skewed, with the largest volumes held by a few of the largest companies. For these reasons, EIA's weekly samples were designed based on the predictive model approach, which leads to the selection of the largest volume facilities as survey respondents. Because weekly surveys cover more than one product and reporting facilities may have locations in more than one region, a multiple-item cut-off sampling procedure was developed. The sampling procedure selected the smallest number of respondents which accounted for at least 90 percent of the total of each product in each region for which data were to be published, based on the previous year's monthly data.

EIA's weekly surveys collect data on inputs, inventories, and imports of crude oil and production, inventories, and imports of major petroleum products (motor gasoline, distillate fuel oil, residual fuel oil, and jet fuel). These data are published 6 days after the close of the report week in the WPSR. Respondents are asked to provide good estimates on the weekly surveys since accounting records are not available at the time the weekly forms must be filed.

EIA's monthly survey forms collect more extensive data based on company accounting records. These data are published in preliminary form in the PSM 60 days after the close of the report month. Final data, reflecting any necessary corrections, are published in the PSA 6 months after the close of the calendar year.

The numbers of respondents to the monthly and weekly surveys are shown in Table 1. Clearly, since even the monthly surveys have relatively few respondents, these survey systems provide a unique opportunity for the evaluation of the quality of the data.

TABLE 1
NUMBER OF RESPONDENTS TO WEEKLY AND MONTHLY SURVEYS
(in Sept 1985)

| Type of Respondent | Weekly | Monthly |
|---|---|---|
| Refineries | 153 | 254 |
| Bulk Terminals | 71 | 310 |
| Product Pipelines | 50 | 87 |
| Crude Stockholders | 87 | 177 |

## A History of Data Quality Activities

In 1982, the PSD initiated a program to evaluate the quality of its data. A comparison of the time series published at different points in time indicated that there were some systematic differences between values based on the weekly surveys and those based on the monthly surveys. Studies were initiated to identify reasons for these differences and to resolve them.

The weekly surveys were begun in 1979, and the data were first published in 1980. The monthly surveys, on the other hand, have a long history. Their precursors began as early as 1910 in the Bureau of Mines. In the early days of the weekly system, the monthly and the weekly surveys were run as completely independent systems. For this reason, not all of the revisions to the monthly surveys were concurrently reflected in the weekly surveys.

Problems associated with operating the weekly and monthly surveys independently were resolved in January 1983 when all of the PSD surveys became part of the integrated Petroleum Supply Reporting System. Now, all of the surveys are redesigned at the same time, and use a single set of definitions. The system also includes "sample control procedures," which formalize the coordination of any changes in the identification numbers of respondents, and keeps track of the weekly sample coverage from month to month.

Data quality efforts continue, with major diagnostic tools provided by data comparisons: comparisons of published data; comparisons of totals of the data submitted to the two surveys by the respondents to the weekly surveys; and comparisons of the data submitted by individual respondents themselves.

## Evaluation of Error in Published Totals

For any product covered in the weekly surveys, three time series can be constructed:

1. A monthly time series calculated from the weekly data. For inventories, the monthly value is interpolated based on the inventories reported for the 2 weeks that bracket the end of the month. For production and inputs the monthly estimate is the sum of the values reported for the weeks which are entirely in the month, and the values for the beginning and ending weeks prorated based on the number of days of the week which are in the month of interest. The monthly estimates calculated from the weekly data will be referred to as the MFW series (preliminary data, based on a sample).

2. The monthly data as published in the PSM. This series will be referred to as the PSM series. The data are based on a census and accounting records, but the preliminary publication.

3. The monthly data as published in the PSA. This series will be referred to as the PSA series. The data are from the monthly surveys used in item 2 but after corrections have been made.

The comparison of published U. S. totals by product provides an estimate of the overall accuracy and consistency of the weekly and monthly series. Every year when the annual data are finalized for publication in the PSA, the Petroleum Supply Division updates its comparison of the accuracy of different time series. An example of this comparison is provided in Figures 2 and 3. These figures were taken from the article "Timeliness and Accuracy of Petroleum Supply Data" which appeared in the June 1985 issue of the PSM. An article with this title has appeared annually in the PSM since 1982.

Figures 2 and 3 illustrate the range of percent differences between preliminary and final published values. The end points of each bar are the maximum and minimum percent deviation during the course of a calendar year. The line indicates the median percent deviation during that year. The first bar in each group compares the MFW value to the PSA value. The second bar in each group compares the PSM value to the PSA value.

Under the assumption that the PSA value represents "truth," these charts permit an assessment of the accuracy of the preliminary published data. In all cases, the range of differences between the preliminary and final monthly published values is smaller than the range of differences between the MFW and the final monthly values. Reasons for differences between PSM and PSA, and for differences between PSA and MFW are described separately below.

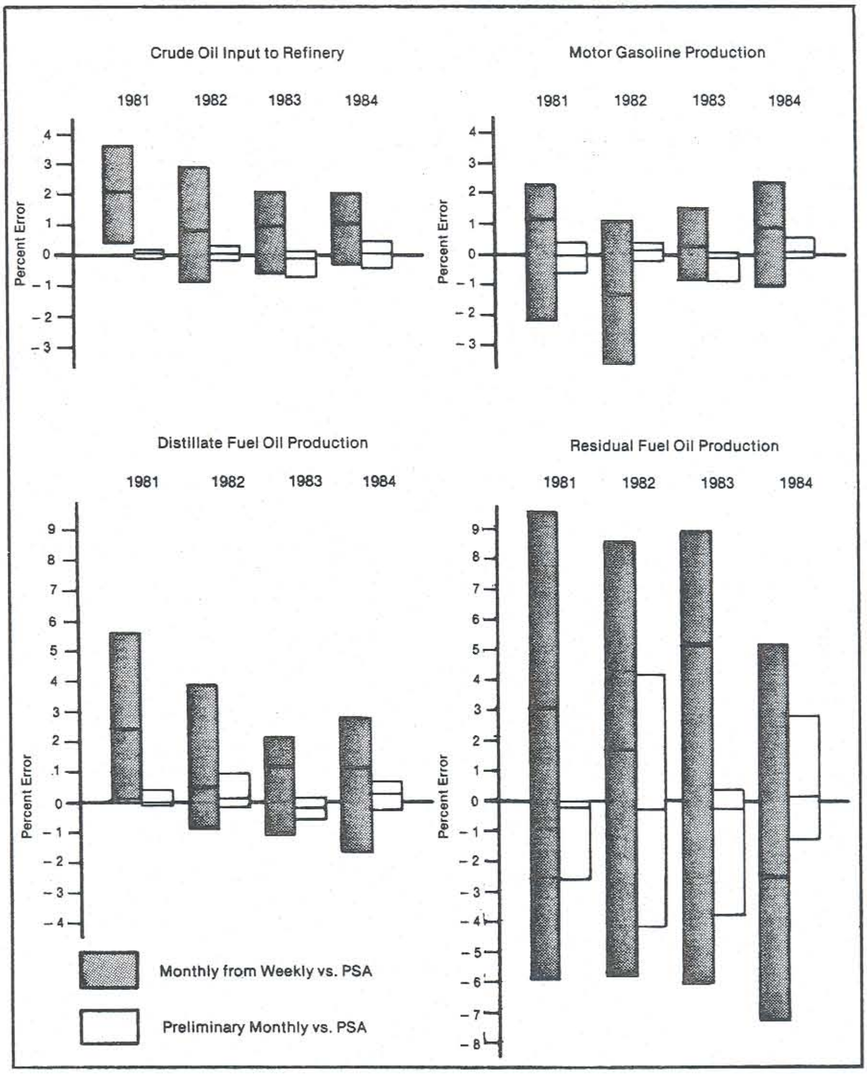
## Differences Between PSM and PSA

Figures 2 and 3 show that there is no systematic difference between the PSM value and the PSA value. Additionally, except for residual fuel oil production, the difference between the preliminary PSM value and the final PSA value is usually less than 1 percent in absolute value. This is expected because the difference between the PSM and PSA values is due only to imputation for nonresponse at the time of first publication, or to errors which were unresolved at the time of first publication.

The PSM surveys are mandatory surveys. On average, at the time of first publication the response rate to the monthly surveys is 99 percent. Additionally, any nonrespondents tend to be the smaller facilities because the extensive telephone follow-up efforts are directed at the larger-volume nonrespondents. The values imputed are the values reported by the nonresponding company in the previous month.

The largest difference between PSM and PSA values for residual fuel oil production was 2.86 percent and occurred in July 1984. It was due to resubmissions by two companies and the correction of a keypunch error for a third. Companies are required to resubmit data to the monthly surveys if they discover that there was a change of more than 5 percent from their initial submission.

Figure 2. Range of Percent Errors for Interim Refinery Inputs and Production Data

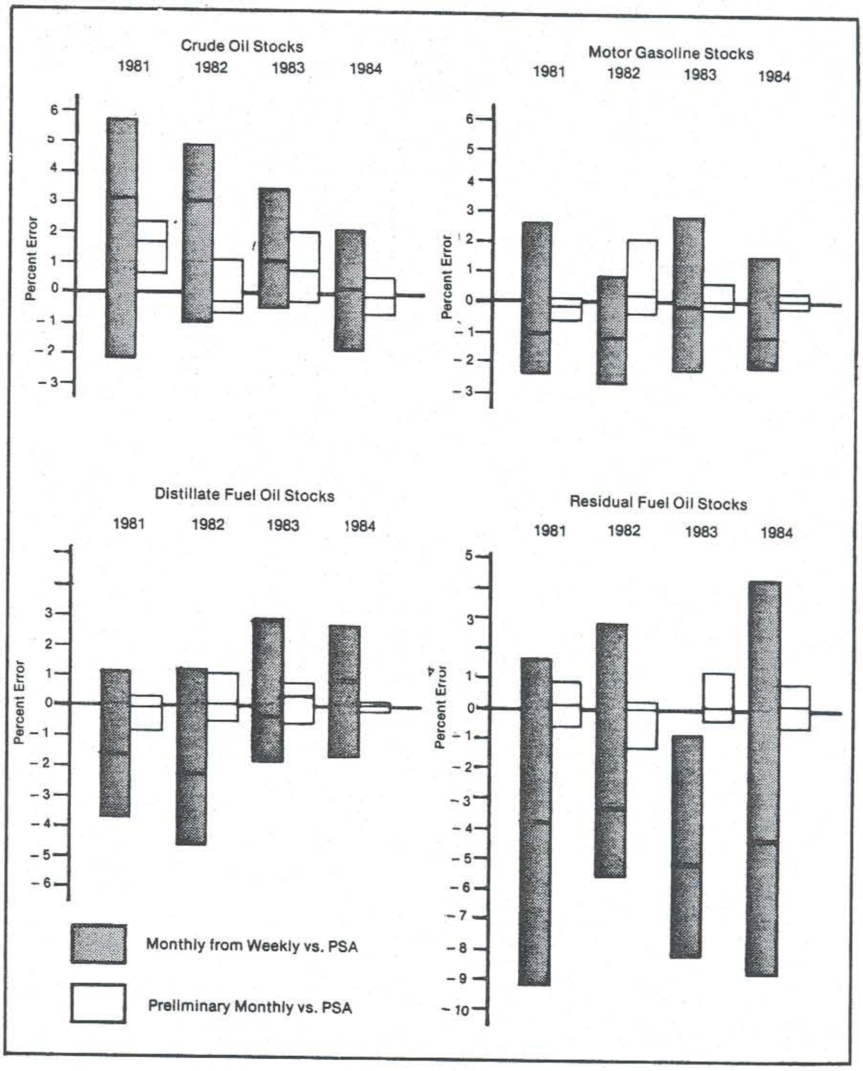


Note: Line = Median of percent errors; i.e., the average of the two middle values when the values are arranged in order of magnitude.

Bar = Range of percent errors occurring during the year; i.e., the upper point of the bar is the maximum percent error and its lower end point is the minimum percent error.

Source: Energy Information Administration, Petroleum Supply Reporting System

Figure 3. Range of Percent Errors for Interim Stocks Data



Note: Line = Median of percent errors; i.e., the average of the two middle values when the values are arranged in order of magnitude.
Bar = Range of percent errors occurring during the year; i.e., the upper point of the bar is the maximum percent error and its lower end point is the minimum percent error.

Source: Energy Information Administration, Petroleum Supply Reporting System

<u>Differences Between MFW and PSA</u>

The first bar in each chart, showing the comparison of the MFW values and the PSA values are longer than the bars comparing the PSM and PSA, indicating that there is more error in the weekly data.

Figure 2 shows that for most of the inputs and production series, the weekly data tend to be higher than the monthly data. This systematic difference had a median value of about 1 percent for crude inputs to refineries, motor gasoline production and distillate fuel oil production. For these series, the maximum and minimum differences were reasonably small, less than 2 percent in absolute value. The residual fuel oil production data is much less accurate. The differences between MFW and PSA ranged from -7 percent to +5 percent in 1984.

Figure 3 shows that weekly crude oil stocks had a positive bias of about 3 percent in 1981 and 1982. It fell to 1 percent in 1983, and in 1984 there was no appreciable bias. For crude oil, motor gasoline, and distillate fuel oil inventories the overall range of differences was about plus or minus 2 percent in 1984. The residual fuel oil inventory data, on the other hand, show a continuing systematic underestimation by the weekly data of 4 to 5 percent, with a large range of errors (-9 percent to +4 percent). This series will be used to illustrate the data quality activities which were implemented in the PSD to address and resolve systematic differences between weekly and monthly data.

## Reasons for Differences

Differences between MFW and PSA data can be attributed to any of the following:

Sampling error

1. Error in estimating the total from a sample.

Nonsampling error

2. Imputation for nonresponse in the WPSR.

3. Data processing errors in the WPSR.

4. Interpolation errors in calculating a monthly value from the weekly data.

5. Respondent error.

Differences between monthly and weekly totals are expected because respondents themselves must estimate their weekly data, and because the weekly totals are estimated from a sample. Additionally, although the companies which report on the weekly surveys also report on the monthly surveys, different people within the company are responsible for preparing the weekly and monthly data. Generally, operational people prepare the weekly data, while accounting people prepare the monthly data. Thus, differences between monthly and weekly data at the company level (respondent error) can arise due to different interpretation of definitions.

The weekly processing system has an extensive edit system which detects significant departures from previous submissions from the same company. Thus, most major departures from past trends (due to data processing error or one-time respondent error) are detected and corrected prior to publication. Additionally, the nonresponse rate for the weekly surveys averages about 2 percent, with the nonrespondents tending to be the smaller volume facilities. Since imputation is based on exponentially smoothed averages of the past responses of the missing companies, the imputation for nonresponse and unresolved errors in data should not be contributors to systematic differences between the two series. Thus, points 2 and 3 above are unlikely explanations for the systematic differences between MFW and PSA data.

Calculation of a monthly value from weekly data introduces some error in the resulting monthly estimate because it is based on the assumption that every day in the week that overlaps 2 months has the same average daily stock change, and the same average daily production. However, it is unlikely that the interpolation process introduces systematic errors.

The sampling and estimation for the weekly surveys are based on Royal's model based sampling approach. Extensive studies showed that the model holds quite well. Additionally, 90 percent of the published weekly value is actual respondent data. The ratio estimate is required only to contribute about 10 percent to the final total. Thus, it is unlikely that sampling error is a contributor to the systematic differences between MFW and PSA.

Thus, the most likely explanation for the systematic difference between MFW and monthly data is respondent error: the systematic reporting of different values to the two surveys.

Because of the high level of coverage of the weekly surveys, and the fact that the same data are reported by all facilities on the monthly surveys, a comparison of the MFW, PSM, and PSA time series based on data submitted to the two surveys by the facilities reporting to the weekly surveys, permits a more detailed evaluation of sources of error. Differences between these time series may be attributable to any of the error sources mentioned above, except the sampling error.

## Reasons for Systematic Differences

The comparison of the MFW and PSM totals by survey for the facilities in the weekly sample showed that for some products the systematic difference between published weekly and monthly totals was due to company reporting differences. A comparison of data submitted to the two surveys by individual respondents identified those facilities contributing most to the systematic error.

This led to the "data discrepancy program," a program to contact those respondents showing the largest differences in order to resolve the reporting problems. Respondents were contacted by mail and by telephone, identifying all of the products which showed significant systematic differences between the values they report monthly and the values they report weekly.

There is a long delay between the time EIA observes systematic differences in company reporting, and the time those companies change their reporting practices. First EIA must detect that the reporting problem exists and is systematic over time, then prepare a formal letter to the company. This is followed by a telephone call. After the calls are made there is a further delay while companies identify the reasons for the discrepancies and resolve them. The program is beginning to show results in the reduction of the systematic differences between the MFW and monthly data.

As an example of the search for reasons for systematic errors, consider residual fuel oil inventories which show a systematic 5 percent underestimation by the weekly from 1981 through 1984. Residual fuel oil inventories are reported on the refinery survey and on the bulk terminal survey. Table 2 shows the difference between the MFW and the PSM totals as reported by respondents in the weekly system in 1984. These differences are due exclusively to company reporting differences.

TABLE 2
DIFFERENCES IN TOTALS REPORTED TO WEEKLY AND MONTHLY 1984
RESIDUAL FUEL OIL INVENTORIES
(percent of PSA total)

| DATE | REFINERIES | BULK TERMINALS |
|---|---|---|
| Jan 84 | -1.00 | -6.13 |
| Feb 84 | -1.19 | -3.65 |
| Mar 84 | .21 | -3.23 |
| Apr 84 | -1.07 | -5.87 |
| May 84 | .06 | -4.38 |
| Jun 84 | 1.33 | -5.62 |
| Jul 84 | - .16 | -5.08 |
| Aug 84 | 1.64 | -3.18 |
| Sep 84 | 1.28 | -3.31 |
| Oct 84 | - .39 | -1.67 |
| Nov 84 | .84 | - .04 |
| Dec 84 | - .09 | -1.32 |

This table shows that there was a clear systematic difference in the reports filed to the bulk terminal survey in the first part of 1984. In PSD's 1984 internal documentation of their data quality activities, five companies were identified as contributing to these discrepancies. All were subsequently contacted as part of the data discrepancy program. Reasons given by the companies for differences between monthly and weekly data were that individuals filling out the forms included more products as residual fuel oil on the monthly than they did on the weekly. Table 2 shows that the systematic error may have been corrected by the last 3 months in 1984. Table 3 shows the differences between the MFW and the PSM values for the first 9 months of 1985.

TABLE 3
DIFFERENCES IN TOTALS REPORTED TO WEEKLY AND MONTHLY 1985
RESIDUAL FUEL OIL INVENTORIES
(percent of PSM)

| DATE | REFINERIES | BULK TERMINALS |
|---|---|---|
| Jan 85 | -2.84 | .97 |
| Feb 85 | - .25 | - .67 |
| Mar 85 | - .15 | - .56 |
| Apr 85 | - .70 | -1.72 |
| May 85 | .87 | .39 |
| Jun 85 | .67 | .20 |
| Jul 85 | - .69 | -1.23 |
| Aug 85 | 2.01 | 1.05 |
| Sep 85 | 2.82 | - .37 |

Even though this table still shows that at times there is a considerable difference between the MFW and the PSM values, there no longer appears to be a systematic difference. The conclusion is that the data discrepancy program has eliminated the systematic difference between the weekly and monthly series, at least for residual fuel oil inventories.

### Summary

The data quality efforts of the PSD during the past few years has had an impact on the overall quality of the published data. The focus of these data quality activities has been to identify and correct problems which are fairly easily identified given the available data. Future work will continue the data quality efforts which have begun, and to implement "performance statistics" which will assist in monitoring the accuracy of published data.

## THE STATE-OF-THE-DATA REPORTS OF THE OFFICE OF STATISTICAL STANDARDS

One quality assurance activity of the Office of Statistical Standards (OSS) is to prepare and publish assessment reports (known as State-of-the-Data reports) that compare published EIA series with other similar series, describe how the series are obtained, and what is known about the quality of the data.

While there are limitations to the comparative approach (it is not always possible to find a fully comparable series that is also reliable), the approach has been useful in raising questions about specific features of a data collection and in identifying errors. As fewer comparative series have been available in recent years, OSS has used another approach as well -- to examine the historical data for anomalies. The following sections present examples of both approaches. The comparative approach is illustrated with data on petroleum, coal, and uranium. The approach based on examining the data for internal consistency is illustrated using estimates of energy consumption.

### Comparative Approach

Many of EIA's data series are based on censuses. Therefore, differences between these series and comparative series in a given year are not due to sampling error but may be due to reporting or processing error, differences in target populations, or differences in the concepts used to measure the item of interest.

## Motor Gasoline Product Supplied

The data that illustrate both the best and worst features of the comparative approach are the data on volumes of motor gasoline available for domestic consumption. The EIA reference series is a proxy for demand called "product supplied." It is the disappearance of motor gasoline from the primary supply system (illustrated in Figure 1) and is calculated as production plus stock change plus net imports. As shown in Figure 4, all comparative estimates had increased relative to the EIA reference estimate from 1977 to 1980. While all of the comparative series have limitations (in fact, the two EIA comparative series have been discontinued because of frame deficiencies), the pattern led to the investigation of the EIA reference series. It was determined that the reference estimate missed certain secondary sources of gasoline supply. (Changes in laws resulted in production of motor gasoline at places other than refineries, i.e. blending stations.) This illustrates the best feature of the comparative approach, when the investigation of a discrepancy leads to uncovering a problem. To obtain more accurate estimates of motor gasoline production, definitions on the data collection form were changed to reflect better the flow of products at refineries, and blending stations were included on the list of respondents.

These changes did not completely close the gap between the EIA estimates for 1981 and those considered to be the most reliable, the Federal Highway Administration (FHWA) estimates, although they narrowed it. The FHWA and EIA collect different data; the FHWA data measure the volume of gasoline on which tax is paid whereas the EIA data measure the disappearance of motor gasoline from the primary supply system. It is also possible that the FHWA double counts interstate sales. This illustrates the worst feature of the comparative approach: not being able to determine which series is correct. It is possible that both series are correct and differ only because they are measuring different concepts. It is also possible that while the series are measuring somewhat different concepts, the reason that they differ has more to do with the fact that one or both series is in error. In a situation such as this, it is difficult to determine the exact cause of the discrepancy because the respondents differ and therefore cannot be matched.

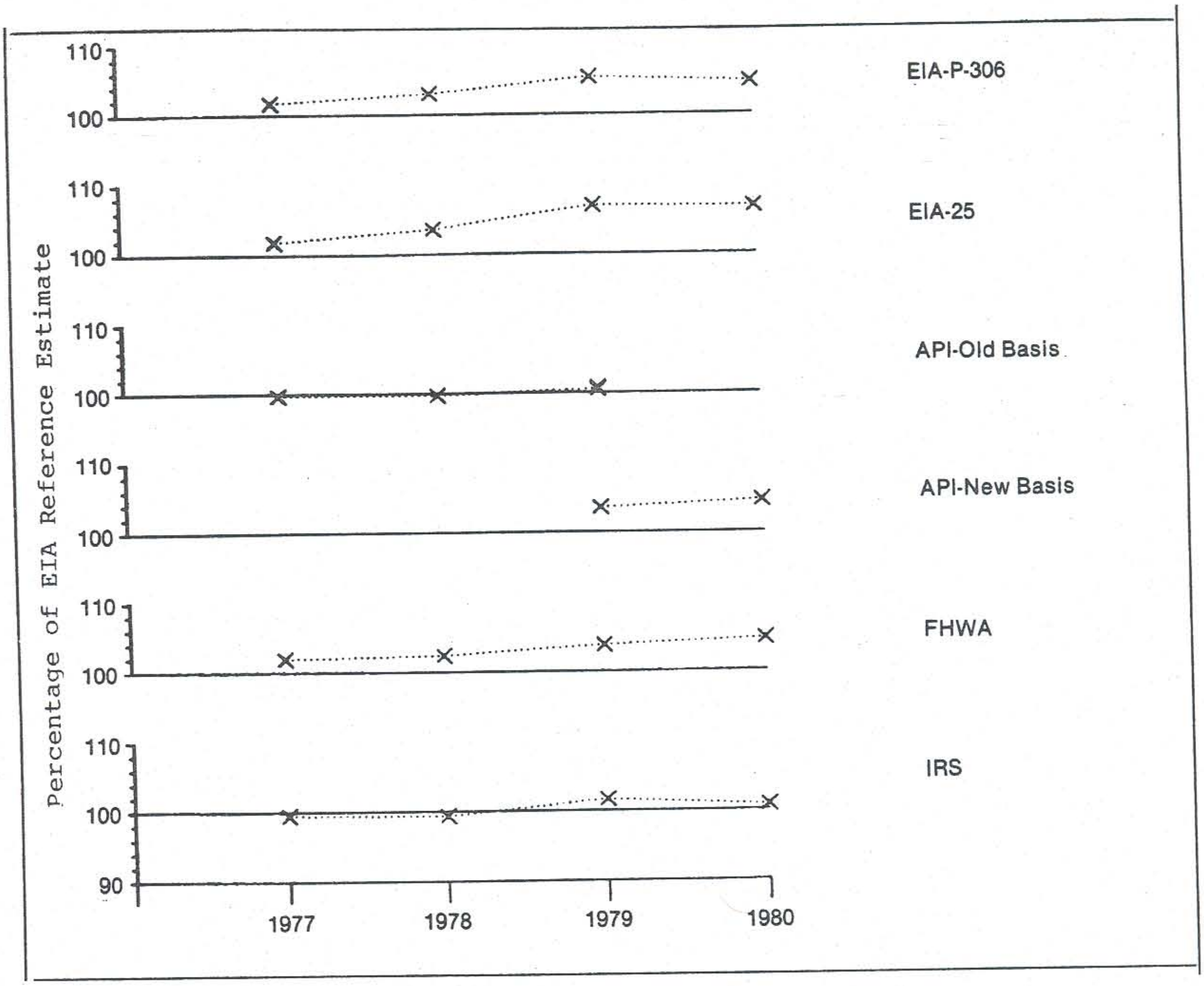
## Coal Production

A series that was ideally suited for the comparative approach was EIA's annual coal production estimates from the Form EIA-7A data collection, "Coal Production Report." These data are aggregated to the county level and published in an annual report, Coal Production. These data were compared with estimates from State mining agencies and were ideal for comparison because the respondents were the same (mines in both cases, although the mines were not always defined in the same way). Also, the coal production data are highly skewed; generally, most of the production in a State is due to a few large mines. Therefore, discrepancies in estimates were usually due to one or two mines whose responses could be verified.

As shown in Figure 5, a large discrepancy between EIA data and the Wyoming State Mining Agency was observed in the estimates of underground production for Wyoming in 1980. The discrepancy was due to one mining company with two mines. This company included both its surface and underground production in the production figure for the underground mine when reporting to the State. Eliminating the surface production resulted in an underground production figure within 0.1 percent of the EIA-7A figure. This conclusion was reached by comparing data on a mine by mine basis. These comparisons were feasible because there were only 26 mines in Wyoming in 1980.

Fairly large discrepancies were observed in the underground production data for Ohio in 1980 and 1981 (Figure 5). The cause of this discrepancy was ascertained by examining county level data. As shown in Table 4, the discrepancy in the State estimates was due to Meigs and Vinton counties. There were two underground mines in Meigs county and one in Vinton, and they were all part of the same mining company. When contacted to explain the discrepancy, the company officials stated that they were reporting raw coal production to the State and clean coal production (a portion of the impurities have been removed) on Form EIA-7A, as required by Form EIA-7A.

Figure 4. Estimates of Volumes of Motor Gasoline Available for Domestic Consumption, 1977-1980



Note: P-306 is Refiner/Importer Monthly Report of Petroleum Product Distribution
EIA-25 is Prime Suppliers System
API is American Petroleum Institute
FHWA is Federal Highway Administration
IRS is Internal Revenue Service

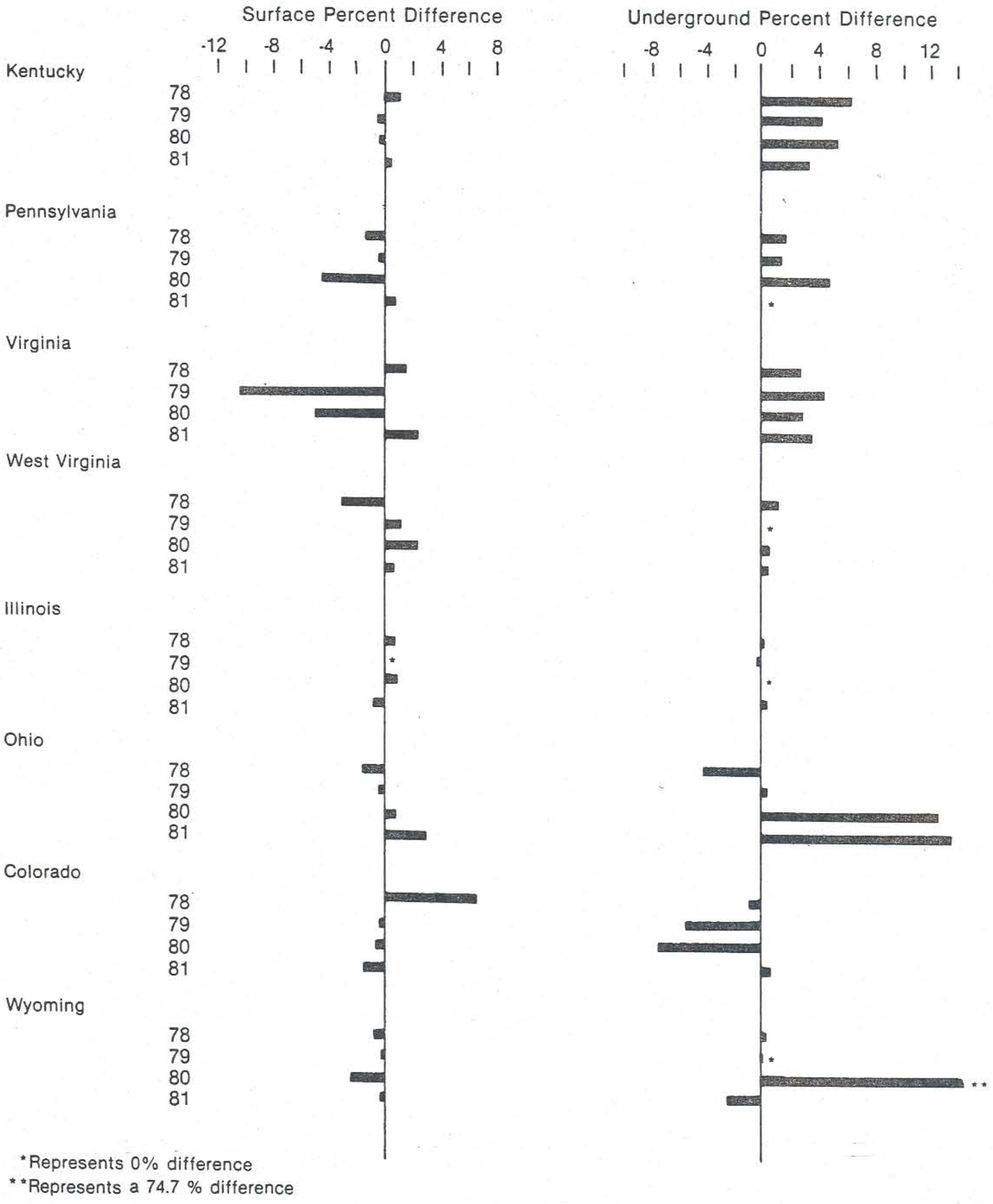
Source: An Assessment of the Quality of Principal Data Series of the Energy Information Administration, April 1983

In contrast to the two previous situations, the discrepancy in the Colorado underground figures for 1979 and 1980 (shown in Figure 5) was due to problems in the EIA data, primarily estimates of production in Las Animas County. Large discrepancies (over 50 percent) were noted for two mines in this county. The two mines, which were part of the same company, were erroneously reporting raw coal data on Form EIA-7A. By 1981 this problem had been corrected, and clean coal production was reported.

Uranium

In contrast to the coal and motor gasoline data, the uranium data posed a different problem, finding comparable series. One of the series of interest was unfilled uranium requirements, that is utilities' uranium requirements that are not covered by usage of inventory or supply contracts as of January 1 of the survey year. Three sources that appeared to have similar data were: Nuclear Assurance Corporation (NAC), Colorado Nuclear Corporation (CNC), and NUEXCO (formerly the Nuclear Exchange Corporation). As shown in Table 5, the estimates for the far future are closer together than for the near future. To better understand the problem, the elements of the definition of unfilled requirements were examined in more detail (Figure 6). While the cause of the difference among the series could not be pinpointed, two areas were found where the instructions on the EIA data collection form were not specific: the inclusion of optional commitments and how to handle cancelled power plants. These items were clarified when the data collection form was revised.

Figure 5. State Mining Agency Estimates of Surface and Underground Coal Production as a Percent Difference of EIA-7A Estimates, 1978-1981



Source: An Assessment of the Quality Selected EIA Data Series: Coal and Electric Power Data from 1977 to 1982, April 1984

## TABLE 4
## EIA-7A AND OHIO MINING AGENCY UNDERGROUND COAL PRODUCTION, BY COUNTY, 1979-1981

(Thousand Short Tons)

| Year | County | EIA-7A (1) | State Reports (2) | (2) As a Percentage of (1) |
|---|---|---|---|---|
| 1979 | Belmont | 4,727 | 4,736 | 100.2 |
| | Carroll | 0 | 3 | -- |
| | Columbiana | 0 | 9 | -- |
| | Harrison | 2,225 | 2,241 | 100.7 |
| | Jackson | 0 | 18 | -- |
| | Meigs | 2,771 | 2,771 | 100.0 |
| | Monroe | 1,823 | 1,828 | 100.3 |
| | Perry | 1,854 | 1,854 | 100.0 |
| | Vinton | 1,087 | 1,087 | 100.0 |
| 1980 | Belmont | 3,374 | 3,291 | 97.5 |
| | Carroll | 0 | 1 | -- |
| | Columbiana | 0 | 8 | -- |
| | Harrison | 1,116 | 1,120 | 100.4 |
| | Jackson | 153 | 153 | 100.0 |
| | Meigs | 3,031 | 4,210 | 138.9 |
| | Monroe | 2,419 | 2,422 | 100.1 |
| | Perry | 1,812 | 1,847 | 102.0 |
| | Vinton | 1,034 | 1,521 | 147.1 |
| 1981 | Belmont | 2,347 | 2,352 | 100.2 |
| | Columbiana | 0 | 9 | -- |
| | Harrison | 1,454 | 1,454 | 100.0 |
| | Jackson | 145 | 152 | 104.8 |
| | Meigs | 2,558 | 3,592 | 140.4 |
| | Monroe | 1,844 | 1,849 | 100.3 |
| | Perry | 1,478 | 1,477 | 99.9 |
| | Vinton | 827 | 1,216 | 147.1 |

Sources: •EIA-7A: Coal Production 1979-1981, Table 4. •State: "Ohio State Coal Report," Table 3, provided by the Ohio Department of Natural Resources, Division of Mines.

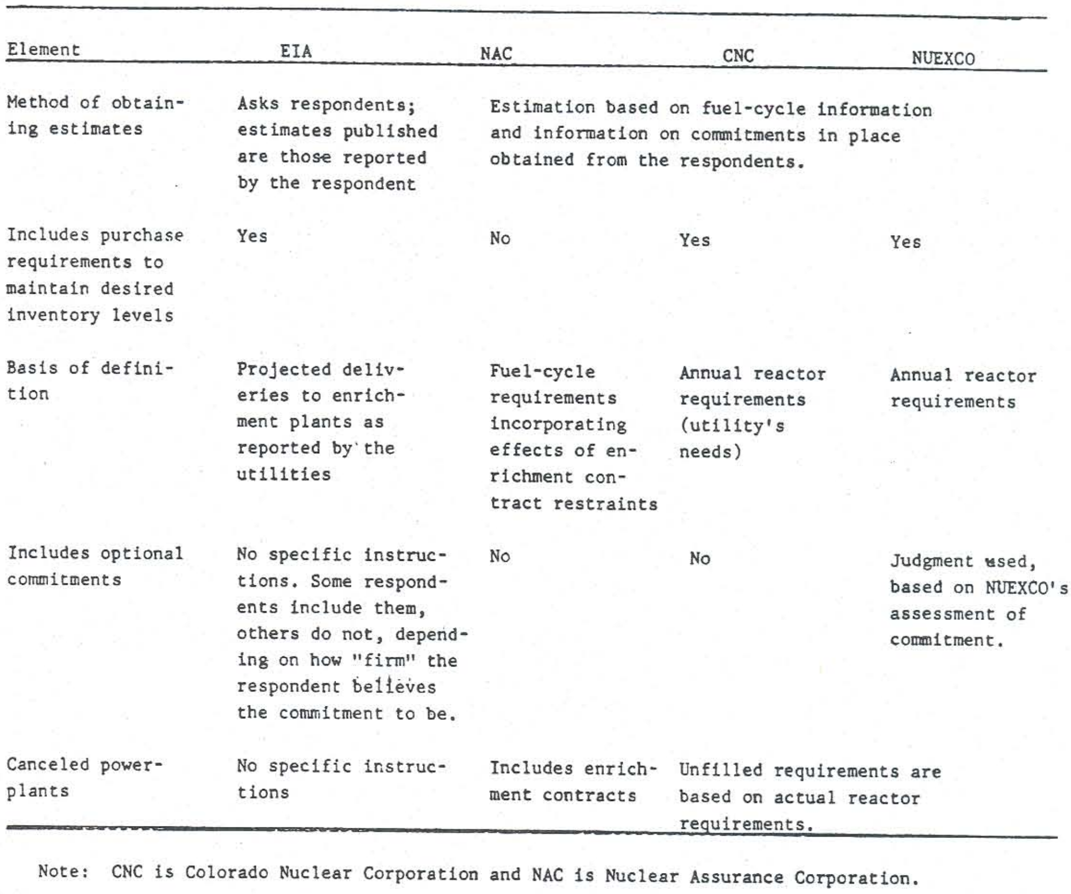
## TABLE 5
## ESTIMATES OF UNFILLED URANIUM REQUIREMENTS AS OF JANUARY 1984

(million pounds)

| Year | EIA | NAC | CNC | NUEXCO |
|---|---|---|---|---|
| 1984 | 2.6 | 0.6 | 0.3 | 1.8 |
| 1985 | 2.3 | 9.1 | 1.1 | 5.2 |
| 1986 | 3.2 | 13.8 | 1.9 | 5.5 |
| 1987 | 5.3 | 16.5 | 5.8 | 8.2 |
| 1988 | 7.6 | 15.5 | 10.3 | 9.7 |
| 1989 | 14.2 | 22.9 | 12.1 | 15.9 |
| 1990 | 12.1 | 20.2 | 17.6 | 21.9 |
| 1991 | 21.6 | 21.1 | 23.0 | 20.0 |
| 1992 | 22.8 | 28.9 | 26.3 | 24.7 |
| 1993 | 28.3 | 32.4 | 29.5 | NA |

Sources:
o EIA: Table 26 of Survey of United States Uranium Marketing Activity 1983.
o NAC: Section F2 of the January 1984 edition of U308 Status Report
o CNC: Special tabulations from the Colorado Nuclear Corporation.
o NUEXCO: Speech on "South Africa's Role in the Uranium Industry," Atlanta Atomic Industrial Forum Conference, Spring 1984.

Figure 6. Comparison of Definition of Unfilled Requirements: EIA, NAC, CNC, and NUEXCO



| Element | EIA | NAC | CNC | NUEXCO |
|---|---|---|---|---|
| Method of obtaining estimates | Asks respondents; estimates published are those reported by the respondent | Estimation based on fuel-cycle information and information on commitments in place obtained from the respondents. | | |
| Includes purchase requirements to maintain desired inventory levels | Yes | No | Yes | Yes |
| Basis of definition | Projected deliveries to enrichment plants as reported by the utilities | Fuel-cycle requirements incorporating effects of enrichment contract restraints | Annual reactor requirements (utility's needs) | Annual reactor requirements |
| Includes optional commitments | No specific instructions. Some respondents include them, others do not, depending on how "firm" the respondent believes the commitment to be. | No | No | Judgment used, based on NUEXCO's assessment of commitment. |
| Canceled powerplants | No specific instructions | Includes enrichment contracts | Unfilled requirements are based on actual reactor requirements. | |

Note: CNC is Colorado Nuclear Corporation and NAC is Nuclear Assurance Corporation.

Source: An Assessment of the Quality of Selected EIA Data Series: Uranium and Nuclear Power Data from 1977 to 1983, April 1985

## Identification of Outlying Observations

Another way of obtaining information about the quality of an aggregate data series is to examine it for outlying observations. The purpose of this examination was twofold: (1) to identify outlying observations that have occurred in past years and (2) to explore techniques that can be used to identify such observations in future years prior to publication.

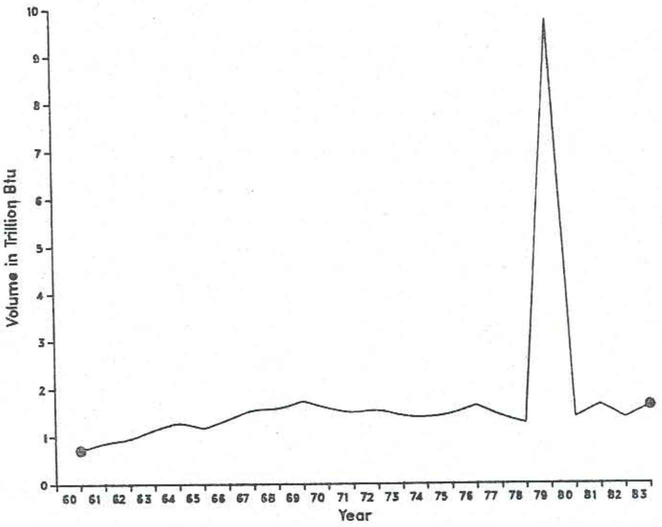
OSS used two approaches to examine EIA's State energy consumption data published in the State Energy Data Report (SEDR). These data are available annually from 1960 to 1983, thus forming a historical series. The State Energy Data System is a composite system. Some of the data are obtained directly from EIA surveys, however, in most cases EIA analysts must make various assumptions and combine data from several sources to obtain estimates needed for the SEDR. While the values are labelled as "consumption," various surrogates are used because State level consumption data are not available. The surrogate measures include deliveries, sales, and product supplied.

One approach used only the time series data and consisted of plotting the data in different ways. The second approach considered other variables as well and consisted of both exploratory and regression analyses. Both approaches had advantages and disadvantages. The first approach was useful in quickly identifying very unusual values. However, unusual values do not necessarily mean that the data are incorrect; such values could arise from unusual weather patterns (e.g. a very severe winter) or changing market conditions. By taking additional variables into consideration, the second approach recognized this possibility. However, this approach was more difficult to implement because additional data had to be obtained.

## Using Time Series Data Only

Historical data were examined both in terms of levels and of year-to-year percent changes. Consumption estimates from 1960 to 1983 were plotted by State, end-use sector (residential, commercial, and industrial), and energy type (e.g. natural gas, electricity, distillate, liquefied petroleum gas). Unusual figures were observed for liquefied petroleum gas (LPG), natural gas and distillate estimates by inspecting these plots. For example, in examining the consumption data on liquefied petroleum gas (LPG) for Delaware (Figure 7), it is obvious that the 1979 figure is an outlier; the volume consumed in 1979 is at least five times higher than the volume consumed for any other year. The LPG sales data used to construct the consumption estimate for Delaware were withheld from publication in 1979 to protect confidentiality. Therefore, the 1979 consumption figure was estimated some other way. Since these data were not withheld in other years, this problem did not reoccur.

Figure 7. LPG Consumption for the Residential Sector in Delaware

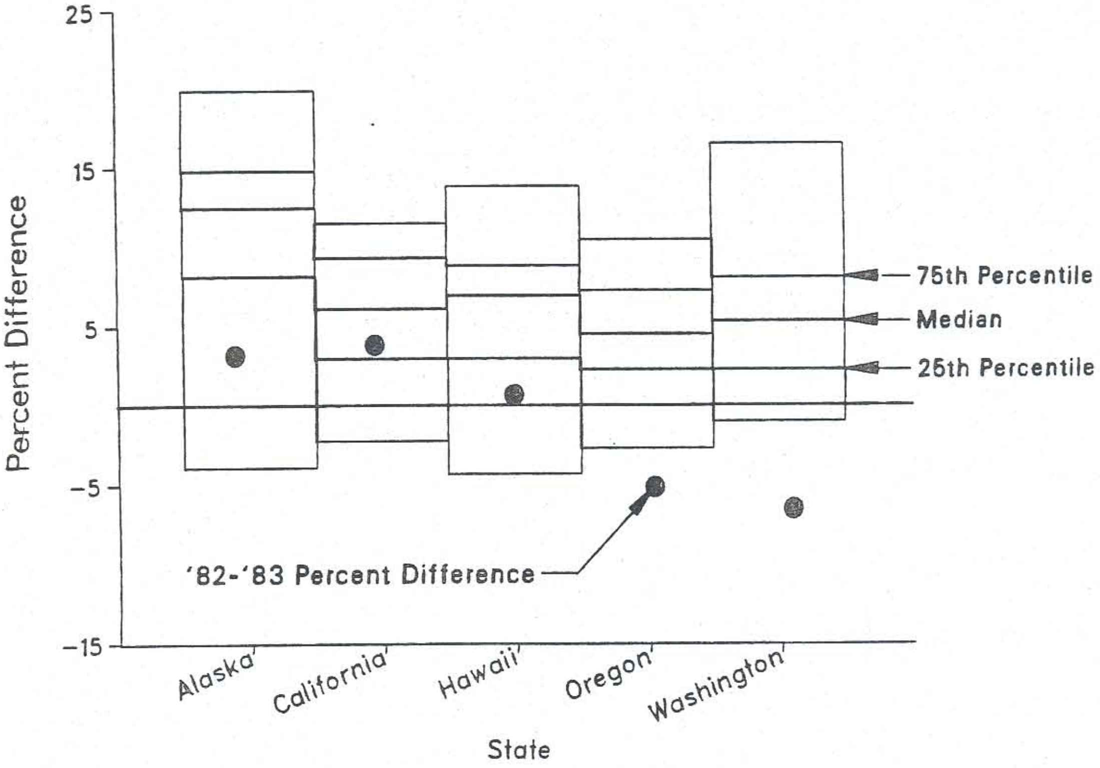


Source: State Energy Data Reports, 1960–1983

Breaks in the electricity consumption series were also observed by examining the plots. For the commercial sector most of the observed breaks occurred between 1974 and 1975 or 1980 and 1981. The pattern was not as clear for the other sectors. Preliminary investigations did not reveal the cause of the breaks during these periods. This series was not investigated further until the regression analysis revealed a similar pattern.

Another graphical procedure that used only the time series data involved comparing the most recent year-to-year percent change in consumption with the changes for the past years. Percentage changes were computed for 22 time periods, beginning with 1960 to 1961 and ending with 1981 to 1982. The percentage change for 1982 to 1983 was then compared with the range for the past 22 years. This procedure was used on data for the residential sector in the State-of-the-Data report. Data on electricity consumption in States in the Pacific Census Division are shown in Figure 8 as an example.

Figure 8. Range in Year-to-Year Percent Differences of Residential Sector Electricity Consumption in the Pacific Census Division, 1960-1982



Source: State Energy Data Report, 1960-1983

As shown in Figure 8, the change in electricity consumption in Oregon and Washington between 1982 and 1983 was outside of the 22-year range. Further investigation indicated that the apparently unusual changes did not represent data problems. Some background on how the electricity consumption figures are obtained is needed before presenting the details. Consumption of electricity is equal to sales. The figures on sales are obtained from a cut-off sample of privately owned electric utilities. In addition, selected publicly-owned electric utilities and other selected utilities such as the Tennessee Valley Authority and the Bonneville Power Administration are included. Expansion factors based on a 1979 census of utilities are used to obtain estimates that represent the entire population of electric utilities.

Examination of the company-level data for Washington and Oregon between 1982 and 1983 showed decreases of similar magnitude for all utilities. Unlike the coal data discussed earlier, the problem was not due to the reporting practice of one company. Furthermore, data from the Edison Electric Institute (EEI) also showed a decrease in residential sales between 1982 and 1983 (Table 6). Moreover, between 1982 and 1983 according to EIA's Electric Power Annual 1983, Seattle experienced a large increase in the price of residential electricity. Among the selected cities for which prices were shown, it was the largest increase during this period. Portland, Oregon also experienced a sizeable increase. These price increases are consistent with the figures showing a decline in sales.

TABLE 6
ELECTRICITY SALES TO THE RESIDENTIAL SECTOR IN
WASHINGTON AND OREGON, 1982 and 1983
(gigawatt hours)

| | EEI | | | EIA | | |
|---|---|---|---|---|---|---|
| | 1983 | 1982 | Percent Diff. | 1983 | 1982 | Percent Diff. |
| Oregon | 13,153 | 14,521 | - 9.4 | 13,116 | 13,825 | -5.1 |
| Washington | 23,905 | 26,633 | -10.2 | 27,266 | 29,157 | -6.5 |

Source:
- o EEI, Statistical Yearbook of the Electric Utility Industry/1983, Tables 41A and 41 B
- o EIA, Electric Power Annual 1984, Table 51.

## Using Additional Variables

An approach that attempts to capture the effect of additional variables on consumption is regression analysis. In the past, EIA staff have discovered problems in the data when using this procedure to study the relationship among variables. The idea was to develop a simple model to predict consumption. The predicted value could then be used to check the estimate of consumption prior to publication. On the one hand, the model had to be simple enough to be implemented as part of ongoing quality control procedures. It would be impractical, for example, to use a model with several equations and a large number of variables for this purpose. On the other hand, the fit had to be good enough so that meaningful comparisons could be made between the predicted value and the estimate for the State Energy Data Report.

The approach was to develop models that included data for all States and all years (cross-section, time series data). There are limitations, however. The use of a single model for all States could ignore important differences. In looking for relationships across series, it is assumed that consumption depends on price, an economic variable such as disposable income per capita, either heating or cooling degree days depending on the energy source, and the value of consumption for the previous time period. An example is presented using data on consumption of electricity in the commercial sector from 1970 to 1982.

The regression model used for the commercial sector is given as:

$$\ln(q_{i,t}) = a + b*\ln(q_{i,t-1}) + d*\ln(y_{i,t}/y_{i,t-1}) + f*\ln(p_{i,t})$$

$$+ g*\ln(c_{i,t}/c_{i,t-1}) + e_{i,t}$$

where i refers to the ith state, t to the tth year, and

$q$ = electricity consumption per capita

$y$ = total disposable income per capita

$p$ = the average price of electricity in the commercial sector

$c$ = cooling degree days

and a, b, d, f, and g are coefficients obtained using ordinary least squares.

While heating degree days and the price of alternative fuels have been included in similar models for other energy sources and end-use sectors, these variables were not significant in this model and were therefore excluded. The regression coefficients are presented in Table 7, and observations with studentized residuals greater than 3 are shown in Table 8. Each of the outliers in Table 8 corresponds to a point where a break appears in the series on electricity consumption in the commercial sector. The consumption estimate changes by at least 20 percent (in either direction) from the previous year for all States shown in Table 8 with the exception of Oregon where the change was 13 percent. There was no corresponding break in the EEI series on sales or on number of customers for these periods.

TABLE 7
REGRESSION COEFFICIENTS AND STANDARD ERRORS FOR COMMERCIAL ELECTRICITY EQUATION

| Variable | Coefficient | Standard Error |
|---|---|---|
| Intercept | 0.253 | 0.023 |
| Consumption (lagged) a/ | 0.934 | 0.008 |
| Disposable Income | 0.264 | 0.059 |
| Average Elec. Price a/ | -0.057 | 0.008 |
| Cooling Degree Days | 0.045 | 0.011 |

N = 624  R-Square = 0.97  Durbin Watson Statistic = 1.86

a/These variables specified as changes.

TABLE 8
OBSERVATIONS WITH STUDENTIZED RESIDUALS GREATER THAN THREE

| State | Year | Residual |
|---|---|---|
| Negative Residuals | | |
| Louisiana | 1975 | -4.54 |
| Nevada | 1979 | -6.15 |
| Nevada | 1980 | -4.32 |
| Oregon | 1981 | -3.14 |
| South Dakota | 1975 | -3.35 |
| Tennessee | 1975 | -3.16 |
| Wyoming | 1975 | -4.68 |
| Positive Residuals | | |
| Alabama | 1981 | 3.46 |
| California | 1975 | 3.34 |
| Iowa | 1981 | 3.04 |
| Missouri | 1976 | 5.51 |
| Wyoming | 1981 | 3.34 |
| Washington | 1972 | 3.88 |

The years that appear most often in Table 8 are 1975 and 1981. There were changes in the expansion factors (used to estimate the total from the cut-off sample) in processing the 1981 data; data from a 1979 census of utilities were used in 1981 whereas data from a 1974 census were used in previous years. It is likely that 1975 was the first year for which the 1974 data were used, but since EIA did not process the data at that point this supposition could not be verified.

## Summary

The comparative approach identified problems due to reporting, frame deficiencies, and definitions of the items that were being measured. The examination for outliers identified changes in the estimation procedures. Attempts have been made by the program offices to resolve the identified problems in the data on motor gasoline supply and uranium. The evaluation of the energy consumption data is very recent and so plans have not as yet been made to resolve identified problems. While not all of the outliers or discrepancies identified by the comparisons resulted from nonsampling error in the EIA data, these methods provided the opportunity to look at the data more closely, and to raise questions, some of which still need answers. Moreover, the methods reinforced findings, as in the case of electricity consumption where comparative data, regression results, and time plots were all available. In retrospect, it may appear that some of the problems could have been identified by studying the data collection form or estimation procedures very carefully. However, the approaches used in the State-of-the-Data report provided the impetus to review definitions and procedures systematically.

## CONCLUSION

This paper has illustrated some of the data quality activities and survey evaluation methods which have been implemented in EIA. The extensive data quality activities of the Petroleum Supply Division provide one example of the types of activities undertaken by EIA's program offices to assure the quality and timeliness of their data. The annual evaluation of selected data series illustrates one aspect of the quality assurance function of the OSS.

In general, overall improvement in EIA data has been a joint effort involving more timely and correct reporting by respondents; the development of efficient processing procedures, improved edit, and estimation methods by EIA's program offices; and more thorough follow-up and reconciliation of aberrant data by the OSS.